Sclera-TransFuse: Fusing Swin Transformer and CNN for Accurate Sclera Segmentation

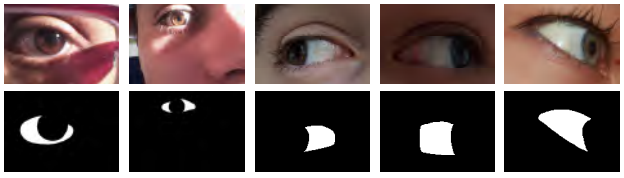
Haiqing Li[1,*], Caiyong Wang[1,†], Guangzhe Zhao[1], Zhaofeng He[2], Yunlong Wang[3], Zhenan Sun[3]

[1]Beijing Key Laboratory of Robot Bionics and Function Research,
Beijing University of Civil Engineering and Architecture, Beijing, P.R. China
[2]Beijing University of Posts and Telecommunications, Beijing, P.R. China
[3]CRIPAC, MAIS, CASIA, Beijing, P.R. China

[*]haiqing_li@stu.bucea.edu.cn, [†]wangcaiyong@bucea.edu.cn(corresponding author)

Abstract

Sclera segmentation is a crucial step in sclera recognition, which has been greatly advanced by Convolutional Neural Networks (CNNs). However, when dealing with non-ideal eye images, many existing CNN-based approaches are still prone to failure. One major reason is that due to the limited range of receptive fields, CNNs are difficult to effectively model global semantic relevance and thus robustly resist noise interference. To solve this problem, this paper proposes a novel two-stream hybrid model, named Sclera-TransFuse, to integrate classical ResNet-34 and recently emerging Swin Transformer encoders. Specially, the self-attentive Swin Transformer has shown a strong ability in capturing long-range spatial dependencies and has a hierarchical structure similar to CNNs. The dual encoders firstly extract coarse- and fine-grained feature representations at hierarchical stages, separately. Then a novel Cross-Domain Fusion (CDF) module based on information interaction and self-attention mechanism is introduced to efficiently fuse the multi-scale features extracted from dual encoders. Finally, the fused features are progressively upsampled and aggregated to predict the sclera masks in the decoder meanwhile deep supervision strategies are employed to learn intermediate feature representations better and faster. Experimental results show that Sclera-TransFuse achieves state-of-the-art performance on various sclera segmentation benchmarks. Additionally, a UBIRIS.v2 subset of 683 eye images with manually labeled sclera masks, and our codes are publicly available to the community through https://github.com/lhqqq/Sclera-TransFuse.



(a) (b) (c) (d) (e)

Figure 1. Examples of challenging eye images with ground-truth sclera masks in non-constrained environments, including (a) eyeglass occlusions, (b) illumination variations, (c) specular reflection, (d) motion blur, (e) gaze deviation.

1. Introduction

For a long time, iris recognition has been the dominant technique of ocular biometrics. However, in recent years, the research community has started to look at other alternative ocular traits besides the iris, *e.g.*, the periocular region [21] and the vasculature of the sclera [15]. Specially, sclera recognition is considered to be highly promising due to several desirable characteristics of the sclera region, *e.g.*, uniqueness, security, stability, user-friendliness [25]. As an initial operation of sclera recognition, sclera segmentation aims to detect valid sclera region containing vasculature structures. This is very crucial, since improper sclera segmentation could either lose important identity-related information or introduce other distractive region, both damaging the accuracy of sclera recognition. However, sclera segmentation is not a trivial task, especially in non-constrained environments, various noise factors in eye images such as blur, occlusions, illumination variations, specular reflections and gaze deviation make it challenging, as illustrated in Figure 1.

Since 2015, a series of sclera segmentation benchmarking competitions in various scenarios, *i.e.*, SSBC 2015 [9], SSRBC 2016 [7], SSERBC 2017 [3], SSBC 2018 [8], SS-

BC 2019 [6], SSBC 2020 [22], and SSRBC 2023 [4] have been held in main biometric conferences (BTAS, ICB, IJCB) and pushed forward the development of sclera segmentation algorithms. Traditional image segmentation algorithms based on shape contour or pixel thresholding were mostly employed in the early competitions whereas recent competitions have been dominated by CNN-based segmentation algorithms, *e.g.*, SegNet [1], U-Net [20], due to their superior segmentation performance. In the literature, Fully Convolutional Network (FCN) and Generative Adversarial Network (GAN) are used for sclera segmentation [17]. Sclera-Net [18], an improved SegNet architecture that exploited identity and non-identity mapping residual skip connections in both encoder and decoder, achieved impressive sclera segmentation performance. ScleraSegNet [25] incorporated channel and spatial attention modules into the central bottleneck part or skip connection part of the original U-Net architecture, which greatly improved the performance of sclera segmentation and won the SSBC 2019 [6]. UNet-P was a modified version of U-Net and ranked first in SSBC 2020 [22], whose core idea was a novel pre-processing procedure to normalize the sclera images before feeding them to the segmentation model.

Although CNN-based methods have emerged for sclera segmentation and are superior to traditional methods in the performance, we find that many existing CNN models are susceptible to various noise and even biased in terms of eye color, ethnicity and acquisition device [23]. These largely result from the fact that convolutional operations, which play a predominant role in CNNs, have limited receptive fields. As a result, it is difficult for CNNs to model global semantic relevance in the image. Additionally, the inductive biases of convolutional architectures may also limit their ability to build global contexts of images. This indicates that CNN-based methods may struggle to provide a personalized guide that can easily accommodate the differences in the target eye image and generate robust sclera segmentation results.

To solve the problem of the lack of receptive fields, self-attention mechanism has been tried to improve the existing models, *e.g.*, ScleraSegNet [25]. However, it is still built upon the convolutional operations. Recently, Vision Transformer (ViT) [11] has been emerging as a pure Transformer architecture in Computer Vision (CV) field, and discarded convolutional operations in its structure. Compared to CNNs, ViT is able to capture long-range spatial dependence through self-attention mechanism applied directly to sequences of image patches, which makes it well-suited for tasks that require modeling global contexts, including sclera segmentation. Despite the potential of ViT in CV field, most Transformer-based works require a large amount of data for training, which limits the application of ViT in the field of sclera segmentation of relatively small-sized datasets. Furthermore, Vit is still not comparable to CNNs in terms of characterizing local spatial details and positional encoding, which are important to refine local sclera segmentation targets.

In this paper, we propose a novel two-stream hybrid model for sclera segmentation, named Sclera-TransFuse. Different from previous sclera segmentation methods, the proposed model enjoys the benefits of both CNNs and vision transformer in the feature extraction. Besides, multiple effective modules and learning strategies are applied into the encoder and decoder of the proposed model, improving the performance of sclera segmentation. The main contributions of this paper are summarized as follows: 1) A two-stream encoder-decoder model is proposed for sclera segmentation, where the encoder integrates CNNs and Transformer in parallel to form complementary feature extractors. The ResNet-34 [12] encoder is used to extract local detail features, while the Swin Transformer [16] encoder is used to model the long-range spatial dependence in the image. Besides, a Swin Transoformer block allows the decoder to refine the skip connected features while capturing global contextual information. 2) A Cross-Domain Fusion (CDF) module is designed to efficiently fuse multi-scale features from CNNs and Transformer through information interaction and self-attention mechanism. 3) Deep supervision strategies are employed to learn intermediate feature representations better and faster. 4) Extensive experiments demonstrate the effectiveness and superiority of our Sclera-TransFuse in the sclera segmentation task. Besides, a UBIRIS.v2 subset of 683 eye images with manually labeled sclera masks, and our codes are publicly available to the research community.

The rest of this paper is organized as follows. Section 2 details the proposed Sclera-TransFuse. Section 3 describes the experimental settings. In Section 4, we present and analyze the experimental results quantitatively and qualitatively. Finally, Section 5 concludes the paper and discusses the future work.

2. Technical details

2.1. Overall Architecture

Figure 2(a) illustrates the overall architecture of our proposed Sclera-TransFuse, which is an encoder-decoder framework similar to U-Net [20]. In the encoder, different from the original U-Net, we design a two-stream hybrid model with two branches for obtaining complementary feature representations. The ResNet-34 [12] with pretrained weights serves as the CNN branch to extract local detail features, while Swin Transformer [16] is chosen as the ViT branch to model the long-range spatial dependence in the image. Hence the RGB eye image $x \in R^{H \times W \times 3}$ is first input into the Swin Transformer encoder and the Resnet-

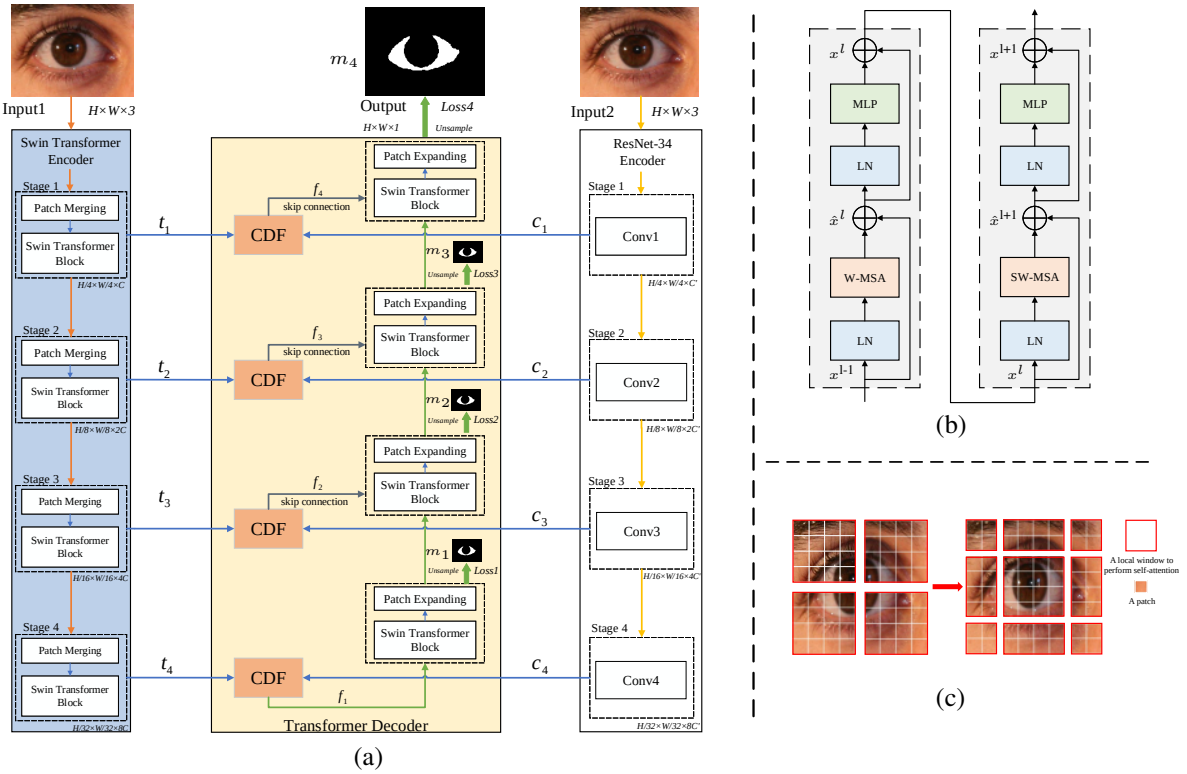


Figure 2. (a) The overall architecture of Sclera-TransFuse model. (b) The detail of Swin Transformer block. (c) An illustration of the shifted window approach for computing self-attention in the Swin Transformer architecture.

34 encoder, respectively, where H represents the height of the image, and W represents the width of the image. Since ResNet-34 and Swin Transformer are hierarchical network architectures, coarse- and fine-grained feature representations are extracted at hierarchical stages and different stages generate feature maps of different sizes. Then a well-designed Cross-Domain Fusion (CDF) module is applied into each encoding stage to fuse the learned two-stream features, yielding a unified feature representation.

As the encoder goes deeper, the size of feature maps gradually decreases. To recover the size of the feature maps and obtain the accurate sclera segmentation results, the decoder adopts a progressive upsampling strategy. To be specific, the fused features at the final stage are firstly processed via a Swin Transformer block (Figure 2(b)) to collect the global contextual information. A patch expanding block is followed to upsample the processed features by a factor of 2. Then from here, the upsampled features from the decoder and the same-sized features from the encoder are concatenated via skip connections. A Swin Transformer block is applied to refine the combined feature and enhance the model's ability to capture global contextual information. Next, a patch expanding block is used to upsample the learned features by a factor of 2. Such procedure is repeated several times until the feature map is restored to half the size of the input image. Finally, a 1×1 convolution is used to adjust the channel number of the feature map to the predicted number of classes (here 1), yielding a 2-D feature map with the size of $(\frac{H}{2} \times \frac{W}{2}) \times 1$. It is further upsampled and reshaped into a 3-D sclera mask $m_4 \in R^{H\times W\times 1}$ as the prediction result.

2.2. The Swin Transformer encoder

The standard ViT [11] architecture requires calculating self-attention between every token and all other tokens in the entire image, yielding a global self-attention mechanism. However, this manner causes an extremely large computational workload and losses the local continuity, which is unaffordable and unfriendly for the dense prediction task (*e.g.*, sclera segmentation). To address this issue, Swin Transformer [16] proposed to limit the self-attention calculation to non-overlapped local windows and enable cross-window connections through the shifted windowing scheme. Specifically, as illustrated in Figure 2(c), the left denotes the window based multi-head self-attention (W-MSA) module using a regular window partitioning strategy, while the right denotes the window based multi-head self-attention (SW-MSA) module using a shifted window partitioning strategy. Both modules are connected in sequence to constitute a Swin Transformer block (Figure 2(b)), which can model the long-range spatial dependence in the image more efficiently.

Mathematically, the Swin Transformer block is formu-

lated as:

$$\begin{aligned}\hat{\mathbf{x}}^{l} &= \text{W-MSA}\left(\text{LN}\left(\mathbf{x}^{l-1}\right)\right) + \mathbf{x}^{l-1},\\ \mathbf{x}^{l} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{x}}^{l}\right)\right) + \hat{\mathbf{x}}^{l},\\ \hat{\mathbf{x}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{x}^{l}\right)\right) + \mathbf{x}^{l},\\ \mathbf{x}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{x}}^{l+1}\right)\right) + \hat{\mathbf{x}}^{l+1},\end{aligned} \tag{1}$$

where $\hat{\mathbf{x}}^{l}$ and $\mathbf{x}^{l}$ denote the output features of the (S)W-MSA module and the MLP module for block l, respectively. LN denotes a LayerNorm layer and MLP denotes a multilayer perceptron network, which consists of two linear transformations with a GELU activation in between.

On the basis of the Swin Transformer block, the Swin Transformer encoder constructs a hierarchical network structure. It consists of four similar stages, each of which contains a patch merging block and several Swin Transformer blocks (here 2 in all stages). Specifically, in the "Stage 1", the patch merging block first splits the input RGB eye image into non-overlapping patches through patch partition. Each patch is considered as a "token" and its feature is the concatenated raw RGB pixel values. Here the patch size is set to 4×4. Then a linear embedding layer is employed to project the original feature of tokens to the C dimension. Swin Transformer blocks are applied to these patch tokens afterwards. In the following three stages ("Stage 2-Stage 4"), the patch merging block is similar to the pooling layer in CNNs. It merges 2×2 neighboring patches into one patch, hence the resolution of the feature maps is reduced by half meanwhile the channel number (*i.e.*, the feature dimension of token) is increased by a factor of 4. Then a linear embedding layer is applied on the merged token features to change the output dimension, followed by Swin Transformer blocks for feature transformation. Overall, four stages generate feature maps $t_1 \in R^{\frac{H}{4}\times\frac{W}{4}\times C}$, $t_2 \in R^{\frac{H}{8}\times\frac{W}{8}\times 2C}$, $t_3 \in R^{\frac{H}{16}\times\frac{W}{16}\times 4C}$, and $t_4 \in R^{\frac{H}{32}\times\frac{W}{32}\times 8C}$, respectively.

2.3. The CNN encoder

The CNN encoder adopts the pre-trained ResNet-34 [12] on ImageNet to extract the local detail features at different scales. We divide the whole architecture into four stages, each of which consists of several convolutional operations with residual connections. Besides, the feature maps from the previous stage are downsampled by a factor of 2 using strided convolution and then proceed to the next stage. Overall, the RGB eye image is input to the ResNet-34 encoder to generate the feature maps $c_1 \in R^{\frac{H}{4}\times\frac{W}{4}\times C'}$, $c_2 \in R^{\frac{H}{8}\times\frac{W}{8}\times 2C'}$, $c_3 \in R^{\frac{H}{16}\times\frac{W}{16}\times 4C'}$, and $c_4 \in R^{\frac{H}{32}\times\frac{W}{32}\times 8C'}$, respectively. Noted that since the subsequent Cross-Domain Fusion module requires the fused features to be consistent in terms of size, we set the channel number of feature maps of the Swin Transformer encoder and the CNN encoder to be equal, *i.e.*, $C = C' = 64$.

2.4. Cross-Domain Fusion Module

As mentioned before, CNNs are good at extracting local detail features, while Transformer is capable of modeling long-range spatial dependence. Thus, they can be used as a pair of complementary feature extractors to capture global contexts while maintaining a strong attention to low-level details. Since there may be domain gaps between the two kinds of features, it is inappropriate directly to concatenate them. To achieve better feature fusion, we propose a Cross-Domain Fusion (CDF) module based on the information interaction and self-attention mechanism, as illustrated in Figure 3.

Specifically, supposed that the feature maps from the Swin Transformer encoder and the CNN encoder in the *i-th* stage are denoted as $t_i \in R^{(h\times w)\times c}$, $c_i \in R^{h\times w\times c}$, respectively. Here $(\cdot)$ indicates the patches are stretched into a 1-D token sequence in calculating the self-attention of Transformer.

To begin with, we use Global Average Pooling (GAP) to compress the two feature maps into $t_i^0 \in R^{(1\times 1)\times c}$ and $c_i^0 \in R^{(1\times 1)\times c}$, respectively. Then we concatenate t_i and c_i^0 along the first dimension for information interaction, resulting in the new Transformer feature map $t_i^1 \in R^{(h\times w+1)\times c}$. Next, t_i^1 is fed into the Swin Transformer block for further enhancing the modeling of long-range dependence and implementing feature refining, which generates the fused feature map $t_i^2 \in R^{(h\times w)\times c}$. For compatibility with the subsequent fusion procedure, the 2-D feature map t_i^2 is reshaped into a 3-D feature map $t_i^3 \in R^{h\times w\times c}$. Similarly, we also obtain the new CNN feature map $c_i^3 \in R^{h\times w\times c}$ after information interaction and a series of processing. Finally, we concatenate the feature maps with different domain semantic properties, *i.e.*, t_i^3, c_i^3, and apply a 1×1 convolution to generate the fused feature map $y_i \in R^{h\times w\times c}$. The above process is named Cross-Domain Interaction (CDI), which is formulated as:

$$\begin{aligned}t_i^0 &= \text{GAP}(t_i), c_i^0 = \text{GAP}(c_i),\\ t_i^1 &= \text{Concat}(t_i, c_i^0), c_i^1 = \text{Concat}(c_i, t_i^0),\\ t_i^2 &= \text{Swin}(t_i^1), c_i^2 = \text{Swin}(c_i^1),\\ t_i^3 &= \text{Reshape}(t_i^2), c_i^3 = \text{Reshape}(c_i^2),\\ y_i &= \text{Conv}(\text{Concat}(t_i^3, c_i^3)),\end{aligned} \tag{2}$$

where Swin denotes the Swin Transformer block.

In addition, since Swin Transformer calculates the self-attention along the spatial axis, we insert a Channel Attention Block (CAB) to promote the learning in the channel dimension as well. Hence we achieve a dual attention mechanism combining both spatial and channel attention. The CAB generates the enhanced feature map $x_i \in R^{h\times w\times c}$, which is formulated as:

$$x_i = \text{Sigmoid}(\text{MLP}(\text{GAP}(t_i))) \otimes t_i, \tag{3}$$

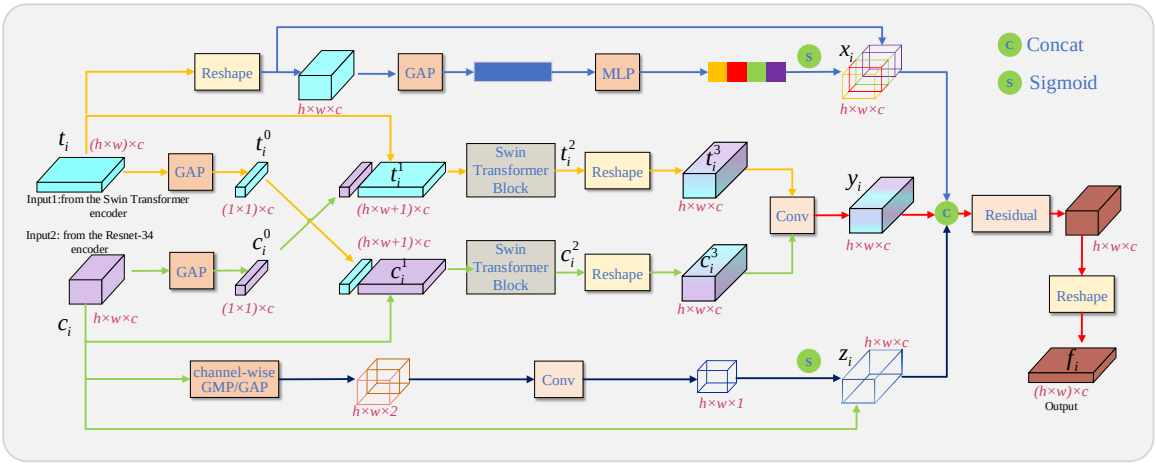


Figure 3. The architecture of Cross-Domain Fusion (CDF) module.

where $\otimes$ represents element-wise multiplication.

For the CNN feature map, we insert a Spatial Attention Block (SAB) in CDF to suppress adverse noise and highlight local details. The SAB generates the enhanced feature map $z_i \in R^{h\times w\times c}$, which is formulated as:

$$z_i = \text{Sigmoid}(\text{Conv}(\text{GAP}_c(c_i) \oplus \text{GMP}_c(c_i))) \otimes c_i, \quad (4)$$

where GAP_c and GMP_c represent global average pooling and global max pooling along the channel axis, respectively. $\oplus$ represents channel-wise concatenation.

Finally, we concatenate the feature map x_i, y_i and z_i, and apply a residual block to refine the fused feature map, which is further reshaped into $f_i \in R^{(h\times w)\times c}$ for decoding.

2.5. Loss function

The proposed Sclera-TransFuse is trained in an end-to-end manner by combining an IoU loss $\mathcal{L}_{IoU}$ and a binary cross entropy loss $\mathcal{L}_{bce}$. Deep supervision [14] has been proven to be effective in helping the model learn intermediate feature representations better and faster. Therefore, similar to the final decoding stage, we predict the sclera masks $m_1 \in R^{H\times W\times 1}$, $m_2 \in R^{H\times W\times 1}$, and $m_3 \in R^{H\times W\times 1}$ at other three decoding stages, respectively. All predicted sclera masks are used for supervision. The overall loss function is formulated as:

$$\begin{aligned}\mathcal{L}_{overall} &= \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4 \\ &= \lambda_1 l(G, m_1) + \lambda_2 l(G, m_2) \\ &+ \lambda_3 l(G, m_3) + \lambda_4 l(G, m_4),\end{aligned} \quad (5)$$

$$l(G, m) = \mathcal{L}_{IoU}(G, m) + \mathcal{L}_{bce}(G, m), \quad (6)$$

where G denotes the ground-truth sclera segmentation mask. λ_1, λ_2, λ_3 and λ_4 are balanced coefficients.

3. Experimental Settings

3.1. Datasets and protocols

To comprehensively evaluate the accuracy and generalization of sclera segmentation models, we adopt both intra-dataset and cross-dataset evaluation settings.

For the intra-dataset evaluation, the sclera segmentation models are trained and tested on the same dataset. The used datasets are listed as follows:

UBIRIS.v2 [19] was originally collected for less-constrained iris recognition and contains 11,102 RGB images from 261 subjects. In [17], a subset of 300 images was manually labeled with sclera masks. Additionally, we also manually labeled a disjoint subset of 683 images to promote this research.

MICHE-I [10] was originally developed for mobile iris recognition, and captured by three mobile devices in uncontrolled VIS conditions. The ground-truth sclera segmentation masks of some images were manually labeled by [17].

SBVPI [24] is a high-quality RGB ocular dataset collected from Caucasian subjects primarily for sclera and periocular recognition research. The sclera segmentation masks were manually labeled by the owner of this dataset.

For the cross-dataset evaluation, following SSBC 2020 [22], the sclera segmentation models are trained on SMD and MASD datasets, then tested on the MOBIUS dataset. The datasets are presented as follows:

SMD [2] was captured by a mobile phone at different times of the day. The dataset has an almost balanced gender distribution but comes with variations in age and skin color of the subjects. The ground-truth sclera segmentation masks were manually labeled by the owner of this dataset.

MASD [5] was captured with a DSLR camera at different times of the day. For each subject, four gaze directions were captured (looking straight, left, right and up) and for

each direction 4 images were taken. The ground-truth sclera segmentation masks were manually labeled by the owner of this dataset.

MOBIUS [22] was captured at four different gaze directions (straight, left, right and up) using three different mobile phones and in three different acquisition conditions. The ground-truth sclera segmentation masks were manually labeled by the owner of this dataset.

More details of these datasets are summarized in Table 1, where the format of eye images is RGB.

Dataset	Reslution	No.of traning	No.of testing	No.of validation
UBIRIS.v2	400×300	400	400	183
MICHE-I	Various	400	400	200
SBVPI	3000×1700	734	734	368
SMD	3200×2400	380	N/A	N/A
MASD	Various	2588	N/A	N/A
MOBIUS	3000×1700	N/A	3494	N/A

Table 1. Summary of the datasets used in this work.

3.2. Evaluation Metrics

Following SSBC 2020 [22], the following evaluation metrics are computed in a pixel-wise comparison manner between the ground truth and the predicted binary sclera mask image:

- **Precision (P)** is the ratio of pixels correctly classified as sclera to all pixels classified as sclera, formulated as $\frac{TP}{TP+FP}$.
- **Recall (R)** is the proportion of pixels correctly classified as sclera w.r.t. all pixels labeled as sclera, formulated as $\frac{TP}{TP+FN}$.
- **F1-score (F1)** is the harmonic mean between precision and recall, formulated as $2\frac{P \times R}{P+R}$.
- **IOU** is the ratio between the size of the intersection of the predicted and ground-truth sclera regions, and the size of the union of the predicted and ground-truth sclera regions, formulated as $\frac{TP}{TP+FN+FP}$.

In the above equations, TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives. Values of the above four metrics are bounded in $[0, 1]$, where the greater value indicates the better segmentation result.

3.3. Implementation Details

The proposed Sclera-TransFuse is implemented in PyTorch and trained on two NVIDIA RTX 3090 GPU of 24GB memory. We adopted the pretrained weights of Swin Transformer and ResNet-34 on ImageNet to initialize the Swin Transformer encoder and the CNN encoder, respectively. During the training, several data augmentation operations such as random rotation, random vertical flipping, and random horizontal flipping are applied. The input images are finally resized to 384×384 pixels using bilinear interpolation for batch processing. The proposed model is trained using the SGD optimizer with a momentum of 0.9 and a weight decay of 0.0005. Moreover, we set the initial learning rate to 0.01, and use a cosine annealing schedule to adjust the learning rate. All models are trained for 50 epochs with batch size of 8, and other similar hyper-parameters are used for fair comparison in the experiment.

4. Experimental results

In this section, we first conducted several ablation studies to analyze the effectiveness of each component used in Sclera-TransFuse. Then, we evaluate the performance of our proposed Sclera-TransFuse in comparison to state-of-the-art methods in the challenging sclera segmentation task.

4.1. Ablation Studies

Influence of encoder. Our proposed Sclera-TransFuse is a two-stream architecture and incorporates two types of encoders, where the Swin Transformer encoder is used to model the long-range spatial dependence while the CNN encoder is used to extract the local detail features. We remove the the Swin Transformer encoder and the CNN encoder, respectively, and compare the remaining single encoder with the original configuration. The results are shown in Table 2. It can be seen from F1 and IOU metrics that both removal operations result in a clear decrease in sclera segmentation performance, indicating that the two encoders are complementary and beneficial for sclera segmentation.

Influence of Cross-Domain Fusion Module. We conduct two kinds of experiments, where the CDF module is replaced with other fusion mechanisms or certain key blocks in the CDF module are removed to assess their effectiveness. For the former, we test a simple fusion strategy termed as "Concat+CNN", *i.e.*, the feature maps from the CNN encoder and the Swin Transformer encoder are concatenated, followed by a 1×1 convolution to adjust the output dimension. For the latter, we remove the CAB, SAB, and both attention blocks (CAB+SAB), respectively. The results are shown in Table 2. From the F1 and IOU metrics, we can see that the proposed model is significantly better than the four fusion methods. Noted that the "Concat+CNN" model is the worst performer among all models, which shows that simply concatenating the information from two encoders is inefficient for cross-domain fusion. Besides, both CAB and SAB are useful in the CDF module. Moreover, the "-CAB+SAB" model only contains the CDI block in the CDF module, but its performance is still superior to the "Con-

cat+CNN" model, which reveals the importance of cross-domain information interaction for feature fusion.

Method	P(%) ↑	R(%) ↑	F1(%) ↑	IOU(%) ↑
Sclera-TransFuse	94.45(03.52)	**94.47(02.60)**	**94.53(01.97)**	**89.69(02.45)**
−ResNet-34 (encoder)	92.74(04.22)	93.70(02.15)	92.84(02.35)	86.72(04.09)
−Swin Transformer (encoder)	92.52(07.35)	93.61(04.86)	92.71(03.98)	86.89(06.40)
△ Concat+CNN (CDF)	92.23(04.42)	94.10(02.50)	93.16(02.24)	87.64(04.66)
− CAB (CDF)	**94.98(04.31)**	93.37(02.02)	94.16(02.32)	88.79(04.15)
− SAB (CDF)	94.40(04.36)	93.56(03.25)	93.98(02.05)	88.03(03.96)
− CAB+SAB (CDF)	92.99(04.83)	94.18(02.06)	93.58(02.30)	87.83(04.58)

− denotes removing certain block.
△ denotes replacing certain block.

Table 2. Ablation experiment of the encoder and Cross-Domain Fusion module on the UBIRIS.v2 dataset. The values in parentheses indicate the standard deviation.

4.2. Influence of deep supervision

The coefficients λ_1, λ_2, λ_3 and λ_4 in Equation (5) are used to balance the contribution of shallow features and deep features in the model, allowing the model to effectively utilize information from multiple levels and mitigate the issue of gradient disappearance. Four different combinations of coefficients are tested, and the results are presented in Table 3. It can be seen that when λ_4 is set to 1 and other coefficients are set to 0.5, the best F1-score and IOU are attained. Besides, when deep supervision is not used ($\lambda_4 = 1, \lambda_1 = \lambda_2 = \lambda_3 = 0$), the sclera segmentation performance is dropped, which indicates the necessity of applying deep supervision.

λ_1	λ_2	λ_3	λ_4	P(%) ↑	R(%) ↑	F1(%) ↑	IOU(%) ↑
1	1	1	1	96.61(02.12)	96.79(03.34)	96.59(02.54)	93.13(03.51)
0.5	0.5	0.5	1	96.59(01.98)	**96.82(03.33)**	**96.66(02.00)**	**93.59(02.85)**
0.1	0.1	0.1	0.7	96.50(02.92)	96.68(04.17)	96.54(02.63)	92.89(03.25)
0	0	0	1	**96.63(02.87)**	96.54(04.06)	96.56(02.86)	92.91(03.11)

Table 3. Ablation experiment of deep supervision on the SBVPI dataset. The values in parentheses indicate the standard deviation.

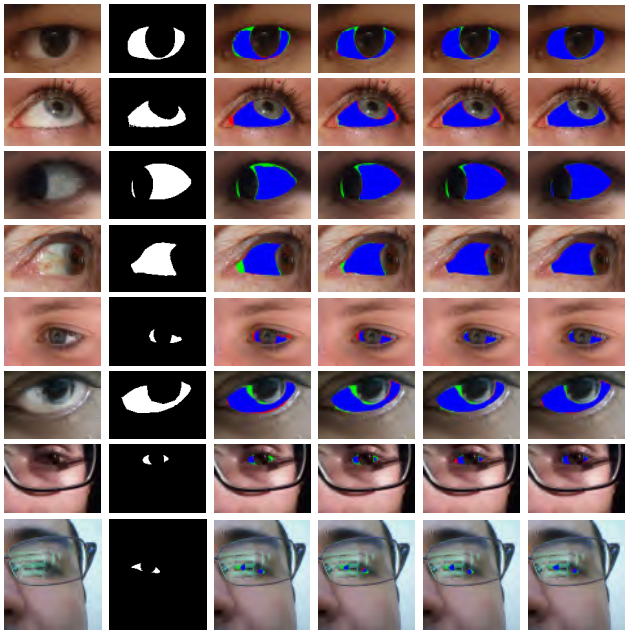
4.3. Comparisons to State-Of-The-Arts

We compare the proposed Sclera-TransFuse with several state-of-the-art methods, including U-Net [20], Sclera-Net [18], nn-UNet [13], ScleraSegNet [25] and UNet-P [22]. The results are shown in Table 4. It can be seen that Sclera-TransFuse achieves the best or second best performance in terms of F1 and IOU metrics under both intra-dataset and cross-dataset settings. Specially, Sclera-TransFuse is competitive with or outperforms ScleraSegNet (the winner of SSBC 2019 [6]) and UNet-P (the winner of SSBC 2020 [22]) in most cases. Figure 4 shows the visualized sclera segmentation results of some challenging samples. It is obvious that our Sclera-TransFuse can predict more accurate sclera segmentation results than other baselines. Overall, both quantitative and qualitative results demonstrate the superiority of our proposed Sclera-TransFuse model.

Training dataset	Testing dataset	Method	P(%) ↑	R(%) ↑	F1(%) ↑	IOU(%) ↑
UBIRIS.v2	UBIRIS.v2	U-Net [20]	91.53(04.96)	90.48(06.58)	90.95(03.16)	83.12(07.90)
		nn-UNet [13]	91.78(02.66)	91.23(04.75)	91.43(02.99)	85.96(02.40)
		ScleraSegNet(SSBC) [25]	91.94(07.27)	91.22(06.86)	91.28(05.56)	84.33(07.38)
		ScleraSegNet(CBAM) [25]	91.74(07.44)	91.24(07.26)	91.20(05.76)	84.21(07.63)
		Sclera-TransFuse	**94.45(03.52)**	**94.47(02.60)**	**94.53(01.97)**	**89.69(02.45)**
MICHE-I	MICHE-I	U-Net [20]	90.60(05.98)	86.05(09.67)	87.83(06.56)	78.85(09.30)
		Sclera-Net [18]	91.88(04.23)	94.69(04.76)	93.13(03.93)	-
		nn-UNet [13]	90.87(03.44)	89.05(04.80)	90.41(04.96)	85.22(03.64)
		ScleraSegNet(SSBC) [25]	89.31(06.12)	90.69(07.34)	89.69(04.88)	81.63(07.49)
		ScleraSegNet(CBAM) [25]	91.71(05.42)	88.11(08.26)	89.54(05.37)	81.45(08.03)
		Sclera-TransFuse	**92.11(03.95)**	**95.69(01.94)**	**93.80(02.22)**	**88.41(03.81)**
SBVPI	SBVPI	U-Net [20]	95.66(02.54)	95.18(04.82)	95.32(02.71)	91.18(04.63)
		Sclera-Net [18]	94.40(03.28)	**98.17(01.60)**	96.24(01.71)	-
		nn-UNet [13]	**96.87(02.96)**	95.12(04.31)	95.67(03.39)	**93.88(02.97)**
		ScleraSegNet(SSBC) [25]	95.39(02.70)	95.86(04.52)	95.53(02.57)	91.55(04.42)
		ScleraSegNet(CBAM) [25]	95.62(02.46)	95.39(04.83)	95.41(02.68)	91.33(04.58)
		Sclera-TransFuse	96.59(01.98)	96.82(03.33)	**96.66(02.00)**	93.59(02.85)
MASD+SMD	MOBIUS	U-Net [20]	95.05(04.69)	70.64(06.87)	80.48(04.30)	77.34(04.66)
		nn-UNet [13]	**95.27(03.44)**	73.89(06.10)	82.13(04.97)	78.55(05.32)
		ScleraSegNet(SSBC) [25]	93.25(06.94)	73.80(05.33)	82.39(07.04)	81.23(06.63)
		UNet-P [22]	90.90(04.00)	83.10(03.00)	86.80(03.00)	**86.80(03.00)**
		Sclera-TransFuse	89.07(10.04)	**86.23(06.81)**	**87.79(07.08)**	85.51(10.52)

− indicates that the value is not available in literature.

Table 4. Comparison of sclera segmentation for different approaches. The values in parentheses indicate the standard deviation.



(a) (b) (c) (d) (e) (f)

Figure 4. Sclera segmentation results of challenging samples. (a) input eye images, (b) ground truth, (c)-(f) segmentation results of U-Net, nn-UNet, ScleraSegNet(SSBC) and our Sclera-TransFuse, respectively. In the predicted results, blue regions represent true positive pixels, whereas red and green regions represent false positive and false negative pixels, respectively.

5. Conclusion and Future work

In this paper, we have highlighted the complementary nature of the CNNs and Swin Transformer, and proposed a novel two-stream encoder-decoder model, namely *Sclera-TransFuse*, for accurate and robust sclera segmentation. A novel Cross-Domain Fusion (CDF) module was introduced to effectively fuse the feature maps from the CNN encoder and the Swin Transformer encoder. Besides, we have innovatively applied the Swin Transformer block

in the decoder to refine the skip connected features while capturing global contextual information. Extensive experiments show the effectiveness and superiority of the proposed Sclera-TransFuse in intra-dataset and cross-dataset settings. For the future work, we will study lightweight Transformer structures to reduce the computational complexity of Transformer-based sclera segmentation models and further apply the segmentation results for sclera recognition. In addition, investigating the usability of Transformer in other biometrics task is also a worthwhile direction to pursue.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62106015), the Pyramid Talent Training Project of BUCEA (JDYC20220819), the Young Elite Scientist Sponsorship Program by BAST (BYESS2023130), and the Beijing Nova Program (Z201100006820050, Z211100002121010).

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] A. Das. *Towards multi-modal sclera and iris biometric recognition with adaptive liveness detection*. Ph. d. dissertation, School of Information and Communication Technology, Griffith University, 2017. 5

[3] A. Das et al. Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition. In *Proc. of IJCB*, pages 742–747. IEEE, 2017. 1

[4] A. Das, A. Mukherjee, U. Pal, P. Peer, and V. Štruc. The 8th sclera segmentation and recognition benchmarking competition (ssrbc 2023). https://sites.google.com/hyderabad.bits-pilani.ac.in/ssrbc2023/home?pli=1. 2

[5] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein. Multi-angle based lively sclera biometrics at a distance. In *Proc. of CIBIM*, pages 22–29. IEEE, 2014. 5

[6] A. Das, U. Pal, M. Blumenstein, C. Wang, Y. He, Y. Zhu, and Z. Sun. Sclera segmentation benchmarking competition in cross-resolution environment. In *Proc. of ICB*. IEEE, 2019. 2, 7

[7] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein. Ssrbc 2016: sclera segmentation and recognition benchmarking competition. In *Proc. of ICB*, pages 1–6. IEEE, 2016. 1

[8] A. Das, U. Pal, M. A. Ferrer, M. M. Blumenstein, D. Stepec, P. Rot, Z. Emersic, P. Peer, and V. Štruc. Ssbc 2018: Sclera segmentation benchmarking competition. In *Proc. of ICB*, pages 303–308. IEEE, 2018. 1

[9] A. Dasa, U. Palb, M. A. Ferrerc, and M. Blumensteina. Ssbc 2015: Sclera segmentation benchmarking competition. In *Proc. of BTAS*, pages 1–6, 2015. 1

[10] M. De Marsico, C. Galdi, M. Nappi, and D. Riccio. Firme: Face and iris recognition for mobile engagement. *Image and Vision Computing*, 32(12):1161–1172, 2014. 5

[11] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*, pages 1–21, 2021. 2, 3

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778, 2016. 2, 4

[13] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 7

[14] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Proc. of AISTATS*, pages 562–570. PMLR, 2015. 5

[15] S. Lee, C. Y. Low, J. Kim, and A. B. J. Teoh. Robust sclera recognition based on a local spherical structure. *Expert Systems with Applications*, 189:116081, 2022. 1

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of ICCV*, pages 10012–10022, 2021. 2, 3

[17] D. R. Lucio, R. Laroca, E. Severo, A. S. Britto, and D. Menotti. Fully convolutional networks and generative adversarial networks applied to sclera segmentation. In *Proc. of BTAS*, pages 1–7. IEEE, 2018. 2, 5

[18] R. A. Naqvi and W.-K. Loh. Sclera-net: Accurate sclera segmentation in various sensor images based on residual encoder and decoder network. *IEEE Access*, 7:98208–98227, 2019. 2, 7

[19] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre. The ubiris. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535, 2009. 5

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of MICCAI*, pages 234–241. Springer, 2015. 2, 7

[21] V. Talreja, N. M. Nasrabadi, and M. C. Valenti. Attribute-based deep periocular recognition: Leveraging soft biometrics to improve periocular recognition. In *Proc. of WACV*, pages 4041–4050, 2022. 1

[22] M. Vitek et al. Ssbc 2020: Sclera segmentation benchmarking competition in the mobile environment. In *Proc. of IJCB*, pages 1–10. IEEE, 2020. 2, 5, 6, 7

[23] M. Vitek et al. Exploring bias in sclera segmentation models: A group evaluation approach. *IEEE Transactions on Information Forensics and Security*, 18:190–205, 2023. 2

[24] M. Vitek, P. Rot, V. Štruc, and P. Peer. A comprehensive investigation into sclera biometrics: a novel dataset and performance study. *Neural Computing and Applications*, 32(24):17941–17955, 2020. 5

[25] C. Wang, Y. Wang, Y. Liu, Z. He, R. He, and Z. Sun. Scleraseget: An attention assisted u-net model for accurate sclera segmentation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1):40–54, 2020. 1, 2, 7